NEWS BOTS
Automating news and information dissemination on Twitter

Tetyana Lokot and **Nicholas Diakopoulos**

So-called "robot" journalism represents a shift towards the automation of journalistic tasks related to news reporting, writing, curation, and even data analysis. In this paper, we consider the extension of robot journalism to the domain of social platforms and study the use of "news bots"—automated accounts that participate in news and information dissemination on social networks. Such bots present an intriguing development opportunity for news organizations and journalists. In particular, we analyze a sample of existing news bot accounts on Twitter to understand how news bots are currently being used and to examine how using automation and algorithms may change the modern media environment. Based on our analysis, we propose a typology of news bots in the form of a design and editorial decision space that can guide designers in defining the intent, utility, and functionality of future bots. The proposed design space highlights the limits of news bots (e.g., automated commentary and opinion, algorithmic transparency and accountability) and areas where news bots may enable innovation, such as niche and local news.

KEYWORDS: algorithmic journalism; automated journalism; bots; computational journalism; news bots; social media; Twitter

Introduction

In early 2015, Automated Insights announced that it was producing and publishing 3000 earnings report articles per fiscal quarter for the Associated Press, all written automatically from structured data via algorithm.[1] Often referred to as "robot" journalism, such technology offers a substantial opportunity for producing online news, not just for writing articles, but also for journalistic tasks such as reporting, curation, or even data analysis and visualization (Broussard 2014; Carlson 2015; Gao et al. 2014; Young and Hermida 2014; Shearer, Basile, and Geiger 2014). Automating the production of news and information offers new possibilities for creating content at scale, personalizing that content given the low cost of adaptation, and covering events more quickly than a human ever could.

As important conveyors of media and other communications, social network platforms represent a domain that is becoming increasingly important and interesting for the study of automated content production and dissemination via bots. Social bots have been defined variously as "automated social actors"—software designed to act

Routledge Taylor & Francis Group
Digital Journalism, 2015
http://dx.doi.org/10.1080/21670811.2015.1081822

similarly to how humans might act in social spaces (Abokhodair, Yoo, and Mcdonald 2015, 840), as "software agents that interact on social networking services" (Hwang, Pearce, and Nanis 2012, 40), and as "automatic or semi-automatic computer programs that mimic humans and/or human behavior" (Wagner et al. 2012, 41). Such bots have been observed across many social networks, like Twitter, Facebook, Reddit, and Wikipedia, interacting with users or content in various ways. A recent report indicated that up to 8.5 percent of accounts on Twitter and about 7 percent of accounts on Facebook may in fact be automated.[2] Many of these bots serve as powerful vectors for spewing spam (Lee, Eoff, and Caverlee 2011) or for manipulating perceptions of political speech (Ratkiewicz et al. 2011), but some also work as positive purveyors of news and information—what we term as a whole "news bots"—automated accounts that participate in news and information dissemination on social networking platforms. Here, we conceive of news broadly, as information that is new, interesting, and somehow relevant to individuals or communities. While classical indicators of newsworthiness may include timeliness, proximity, importance, impact, conflict or controversy, sensationalism, prominence, and novelty or the unusual (Shoemaker, Chang, and Brendlinger 1987), other scholars suggest news is a more "slippery" concept, and provide journalists' own straightforward definition of news as timely information about newsworthy people and events (Hall 1981, 231). In this work we focus our study on the ways in which such news bots are being used on Twitter for constructively sharing such broadly defined news and information. How do these bots change the media sphere and the provision of news information as part of journalistic practices?

In particular, we undertake a qualitative analysis of 60 news bots collected via a purposive sampling technique, and another 300 potential news bots collected via a targeted Twitter sample via DataSift. Our results indicate the myriad ways in which news bots are being employed for topical, niche, and local news, as well as for providing higher-order journalistic functions such as commentary, critique, or even accountability. We provide detailed examples illustrating the different functions of bots that we observed. We contribute a detailed typology and characterization of these bots which (1) creates an editorial decision space that can be a starting point for defining, designing, and building social news bots that play constructive roles on social media, and (2) highlights opportunities for future development and design solutions. Moreover, by more deeply understanding how news bots are being constructively and productively employed in news and information dissemination, we contribute a discussion of how such entities participate in the wider media environment and expose issues of automation and speech related to platforms, transparency, and accountability.

Related Work

Bots are not an entirely new phenomenon and have been observed and studied in a variety of human communication platforms, most recently in social networks (Abokhodair, Yoo, and Mcdonald 2015). The uses and activities of bots that have been observed on social media, our focus in this work, are myriad, including social shaping (Hwang, Pearce, and Nanis 2012), content pollution (Lee, Eoff, and Caverlee 2011), social metric gaming (Cook et al. 2014; Messias et al. 2013) or ranking manipulation (Mustafaraj and Takis Metaxas 2010), infiltration (Boshmaf et al. 2011), political astroturfing (Ratkiewicz et al. 2011) and misdirection (Abokhodair, Yoo, and Mcdonald 2015), recommendation (Aiello et al. 2012), scholarship dissemination (Haustein et al. 2015), and activism or advocacy (Wilkie, Michael, and Plummer-Fernandez 2014). Moreover, a thriving bot-making culture crafts what might be considered digital humanities bots and convenes workshops like the Bot Summit[3] to discuss relevant issues. Such bots can be found highlighting poetic writing (e.g., @pentametron, @HaikuD2), finding humorous word plays (e.g., @portmanteau_bot), editing video stories (e.g., @eventuallybot), and exhibiting other creative applications.

Less research has focused specifically on the realm of news and information bots, though there have been studies of bots that can identify and share newsworthy events (Steiner 2014). Bots and automated information sharing has been observed in natural disaster events (Starbird et al. 2010), as well as other forms of breaking news events (Mittal and Kumaraguru 2014). Such observations demand that we reconsider the media sphere that emerges on social media not just as a result of human communication, but as a confluence of humans interacting with other humans as well as with automation (Larsson and Moe 2014). The present work seeks to make progress in examining the ways in which news bots are specifically being employed to constructively shape the news and information available on Twitter.

The extant research on social bots often focuses on the potentially detrimental effects of automated information sharing. There is a concern about the use of bots to manipulate public opinion (Ratkiewicz et al. 2011), pollute the content stream with spam (Lee, Eoff, and Caverlee 2011), compromise privacy (Boshmaf et al. 2011), or game social signals (Haustein et al. 2015; Hwang, Pearce, and Nanis 2012). There is a worry of human susceptibility and likelihood to be influenced by such automata (Wagner et al. 2012; Wald et al. 2013). Media stories have reported on the use of botnets to drown out dissent in Russia[4] or to manipulate trending topics in Mexican elections.[5] Recent research reports on an in-depth examination of the use of a botnet to manipulate information flows in both English and Arabic related to the Syrian Civil War (Abokhodair, Yoo, and Mcdonald 2015). As a result, many studies focus on the identification and detection of bot accounts via the extraction of features like temporal activity, network structure, and sentiment to develop machine learning classifiers (Chu et al. 2010; Dickerson, Kagan, and Subrahmanian 2014; Ferrara et al. 2014; Tavares and Faisal 2013).

There is no doubt that there are many nefarious uses for bots in online media, and that the development of such classifiers is important work aimed at quickly curtailing abuses. But far less attention has been paid to the potentially positive and beneficial utility of automated news and information sharing on social networks that we study in the present work, including the ways in which bots may contribute to positive effects in the public media sphere if employed ethically and conscientiously. Research has found that a Twitter bot sharing public health information was perceived as a credible source of information, suggesting that such bots could be gainfully employed by news organizations if properly harnessed (Edwards et al. 2014). In comparison to the extant studies that have reported on the use of bots to share news and information (Mittal and Kumaraguru 2014; Starbird et al. 2010), we take a qualitative approach and examine the constructive utility of a wide sample of bots in the public media sphere as it relates to modern journalism.

Analysis of News Bots

As part of our analysis, we describe the methodology and process of data collection and bot identification mechanisms, as well as provide a summary of our findings, including a typology of key identifying traits and notable observations about some of the bot types and their occurrence in our samples.

Data Collection

In order to learn more about news bots and the specific traits they exhibit that could inform our bot typology, we decided to first examine Twitter accounts that were "known" bots. These were either explicitly designated as bots by their creators or discussed as interesting Twitter bots in thematic articles or "listicles" on news websites, or blog posts by Web developers and bot creators. Because we wanted to examine bot accounts in detail, purposive sampling (Jupp 2006) was deemed an adequate selection method. Though this form of sampling is less concerned with statistical generalizability or representativeness, it aims to reflect the diversity within a given population, especially in qualitative analysis (Barbour 2001). We wanted to have a diverse initial sample of bot accounts in order to then extract traits and qualities from them that would inform the criteria used in the collection of a larger complementary sample from the Twitter firehose.

We collected our initial purposive sample of accounts using search engine results from Google for searches of "bot," "twitter bot," "news bot," "automated twitter account," such as articles and blog posts about Twitter bots, as well as the accounts themselves found directly through Twitter Search on keywords like "bot," "twitter bot," and "automated account." Our searches were often triangulated if more than one source mentioned the same account, making them prima candidates for our sample. While sampling, we assumed self-described bots were honest in their self-representation or that if open-source code was published, or if the creator of the bot was acknowledged, that these were automated accounts. After collecting a primary set of accounts, we excised accounts that were not run in English and accounts that were clearly not bots (i.e., they did not conform to the definition of "automatic or semi-automatic computer programs that mimic humans and/or human behavior" [Wagner et al. 2012, 41] or did not exhibit criteria articulated in other bot labeling methods [Chu et al. 2010; Dickerson, Kagan, and Subrahmanian 2014]), which resulted in a primary purposive sample of 60 news bot accounts.

But since 60 is a relatively small number of accounts for analysis, we sought to expand our sample of news bots through automated means. As prior work has sought to identify bots on Twitter (Ferrara et al. 2014), we first considered whether such a classifier could be used to identify possible news bots to augment our sample. To assess the validity of this technique for identifying news bots in particular, we used the Bot or Not service[6] (Ferrara et al. 2014) to determine how our 60 purposive sample bots were classified. Only 18 out of 60 bot accounts ranked above 50 percent in the overall "botness" score computed by Bot or Not. This meant that 42 of the known bots were interpreted by the Bot or Not algorithm as not bots, i.e. as false negatives. The inaccuracy and margin of error (70 percent) with Bot or Not were too high to be able to effectively use it to classify news bot accounts to increase our sample.

In order to devise a more effective mechanism for sampling accounts on Twitter that were likely to be news bots, we manually examined the purposive sample including account titles, handles, and bios to develop a set of textual features that could be used to scan for and collect news bot candidates using the DataSift service, which at the time had access to the Twitter firehose. We also extracted the text from all the bios of the purposive sample and used the AntConc tool (Anthony 2004) to rank unigrams from the bios according to a keyness metric against a reference corpus of bio text extracted from an arbitrary random sample of English Twitter accounts also collected from DataSift. We thus produced a list of most common keywords contained in the account metadata that were indicative of "newsbotness." These keywords/phrases included: "bot," "by @" "news," "github," "robot," "monitoring," "automated," "alerts," "feed," "aggregated."

We applied these keywords as search and filtering terms to the Twitter firehose using DataSift, and collected a corpus of over 35,000 potential Twitter accounts over two hours on December 3, 2014 whose metadata contained one or more of the search terms, indicating they were a news bot candidate.

Analysis Methodology

In order to test our keyword sampling mechanism, we randomly drew a manageable subsample of 300 unique accounts (~1 percent of those collected) from the DataSift sample and examined them carefully to determine whether or not they were in fact bots using criteria articulated in published bot labeling methods (Chu et al. 2010; Dickerson, Kagan, and Subrahmanian 2014). These criteria included tweet syntax metrics such as the presence of hashtags or mentions, account behavior metrics such as tweet frequency and tweet repeats, as well as account properties like bio content and protected or verified status (Chu et al. 2010; Dickerson, Kagan, and Subrahmanian 2014). Overall, out of the 300 Twitter accounts in this sample, we found that a little over one-third of the accounts, 118, were not bots, and four had no tweets or the account had been suspended. Of the 178 accounts which we coded as "bots," 24 were "cyborgs" and contained an apparent combination of automated and human tweeting. The error rate (118 plus 4, or 40 percent of 300 were false positives) was acceptable given the nature of our qualitative study. One of the biggest culprits in generating false positives—accounts that were not bots but were in the sample because they corresponded to our keyword search criteria—seemed to be the expression (by @handle): it was used by 17 out of 60 bots in our purposive sample to designate the creator, but was also used by many non-bot accounts in the larger DataSift sample to indicate the human who was managing, editing, or running the account.

Our final combined set of news bots thus included the 60 from our purposive sample, plus 178 from the DataSift sample, resulting in a final analysis sample of 238 newsbots. Where appropriate in our findings, we contrast the types of bots that were observed in the purposive sample and the DataSift sample. On the combined sample we created observational descriptions of the information available from each account, such as title, handle, bio, and the latest tweets in the account's feed. Accounts were analyzed in the context of the original Twitter webpage to see how they tweeted and interacted with other accounts, and to see how they presented themselves on the service. For each account, we thus created a qualitative description, including observations of characteristics of the accounts and what seemed unique about the bots and their behavior. We then analyzed these descriptions and the accounts through iterative qualitative coding including open coding, affinity diagramming, typologizing, and memoing (Lofland and Lofland 2005) that inform the typology and findings we report next.

Findings

Four key categories emerged during the open coding and typologizing process. These categories reflect both the bots' algorithmic logic and the editorial decisions behind it. The categories include:

Inputs and sources: what a bot takes in in order to produce some kind of output.
Outputs: what a bot produces given the kind of input it receives.
Algorithm: how the bot processes the input to get the output.
Intent/function: why the bot processes the inputs to get outputs in this particular way.

Each news bot in our sample can be typologized using any of the above categories and their subcategories, and this approach allows us to see the differences between our purposive sample and the larger DataSift sample. Such a typology also creates a multidimensional space that allows for a more comprehensive understanding of what news bots can do and how they can be designed to do it.

Inputs and Sources

Inputs are defined as what a bot takes in in order to produce an output. Twitter news bots in the sample have a single source or multiple sources of input data, meaning that the source categories for one bot may include more than one source type. The sources may include *a single news website or blog*, *several news websites or blogs*, *an RSS feed or feeds*, *another social media platform*, *a hashtag* (e.g., the automated account @yegtraffic provides information on traffic and road incidents in Edmonton, and retweets every tweet with the hashtag #yegtraffic), *a database or an archive of stories*, *files or other data*, *other Twitter accounts*, or *content of tweets the bot reacts or replies to* (including both standalone tweets, as in the case of @beExplicit, a bot that reacts to passive voice in tweeted headlines and makes editing suggestions; and replies at the bot, as in the case of @BBCWeatherBot, an account which tweets weather forecasts in reply to users who request a forecast for a particular geographic area).

Sources of input are *explicit* in very few cases, for example when mentioned in the bio (e.g., @mmorpg_tweeter mentions the five websites it draws video game news from, while the @Treasury_io bot points to a particular database of US treasury data from which its tweets are generated); *evident or obvious* in some cases, when sources can be easily determined from the Twitter feed (as in most single-source cases); or *undefinitive*, as in most multiple-source cases where there is no attribution or any direct way to glean the exact array of sources used.

The purposive sample (60 accounts) included 26 cases (43.3 percent) where the source was explicit (i.e., stated explicitly in the bio), 10 cases (16.7 percent) where it was evident from the account information (possible to infer from tweets, but not stated explicitly), and 24 cases (40 percent) where the source was undefinitive. The DataSift sample (178 bot accounts) included only 10 accounts (5.6 percent) with explicit sources, 85 cases (47.8 percent) where the source was evident, and 83 cases (46.7 percent) where the source was undefinitive. The large number of undefinitive cases, and the significant amount of evident, but not explicitly identified sources could have implications for transparency and credibility of automated journalism, which we address further in the discussion.

Among the types of input, single website and multiple website were the most frequently occurring types in both samples. A detailed breakdown of source input types is shown in Figure 1. Note in the figure that the "unclear" type is a subset of "undefinitive" sources, since even if the sources were not definitive, their type (e.g., multiple websites) could sometimes be surmised.

Outputs

The output of a bot is what the bot produces given the input it receives. On Twitter we found these can take the form of *topical feeds* (with news on a given topic from a single or multiple sources): popular topics include *politics*, *business*, *finance*, *disaster and incidents*, *entertainment*, *security*, *technology*, *sports*, *traffic*, and *weather*. Some traditional news beats, like *health* and *science*, have little or no presence in our sample. A *single-source* feed can also be more *general* in nature, when it draws from a

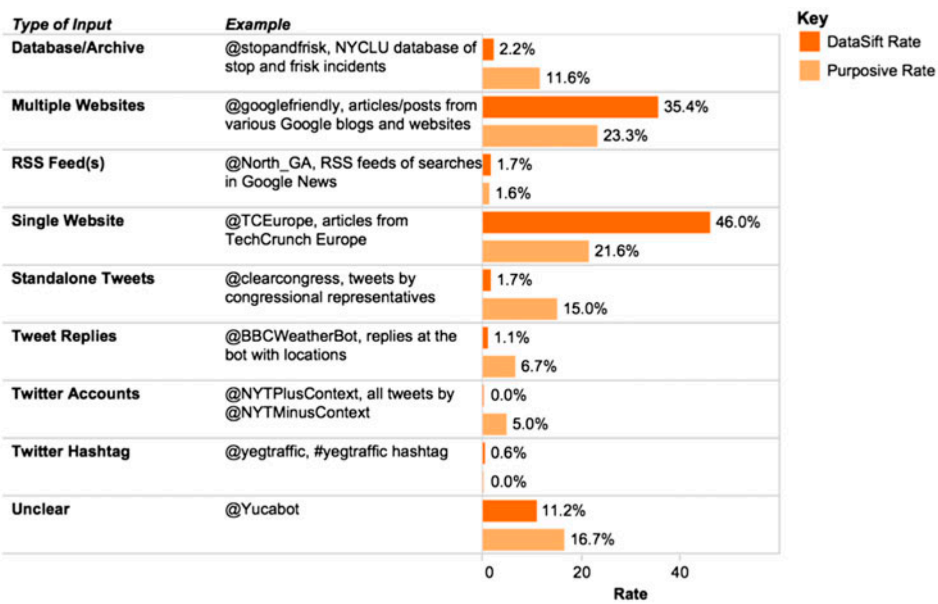

FIGURE 1
Prevalence of types of input source (categories may overlap)

news source that is itself not specialized, but has different kinds of news. Single-source feeds also demonstrated the largest human involvement alongside automated activity: for instance, cyborgs like @washingtonpost or @TheAtlantic which combine general news tweets linking to the main website with retweets and other human content accounted for 20.7 percent (17/82) of the single-source category in the DataSift sample. In other words, 71 percent (17/24) of all of the cyborgs that we observed in the DataSift sample were for single-source feeds.

A *geo-specific* output positions and filters the news and information within a geographic area, whether *national or local*. The main filtering criteria in this case are relevance of a news story to the designated *location* and its *scope*.

Geo-specific feeds, especially those at county or city level, are often traffic or weather updates, in addition to general news. Such feeds can be single source and draw from a local news website (such as @PelhamPatch with news for Pelham) or multiple source (the @North_GA news bot draws from a Netvibes aggregator page with RSS feeds set up for Google News searches of city names from the North Georgia area). Some weather bots, many of which were also geo-specific, rely on input from Twitter users (@BBCWeatherBot asks to tweet at it with a location), while others just tweet local weather for a specified location at regular intervals.

Niche feeds cover a very specific, narrow beat, not typical for traditional journalism. The input is filtered or curated for a small subset of information, relating to a very particular interest or area, such as *survival preparedness* (@BadBluePrep), *MMORPG video game news* (@mmorpg_tweeter), *drone strikes* (@dronestream) by the US Government, *stop and frisk incidents* by the City of New York Police Department in 2011 (@stopandfrisk), or *anonymous wikipedia edits* by IP (Internet Protocol) addresses in the US Congress (@congressedits).

Niche feeds have a very narrow focus, often working within a topic, but providing news on a very specific subset or one segment of it, like a single sports team (e.g., @BuckeyesSpyder). Niche bots can also deal with a specific persona (such as the @shakespearelogs bot, which presents itself as an "experiment to promote blogs and news websites about William Shakespeare via twitter" and curates only Shakespeare-related content) can work with a single input source (@congressedits works with a database of anonymous Wikipedia edits, while @stopandfrisk draws from a data archive of incident reports) or curate content from multiple sources (@BadBluePrep aggregates survival and preparedness stories). In some cases the sources are explicit (@mmorpg_tweeter lists the five websites it curates news from in its bio), in others, like @BadBluePrep, the sources are not evident.

Some news bots do not tweet news directly, but provide *commentary* to certain kinds of information found on Twitter or elsewhere, essentially illustrating a particular point or drawing attention to a particular angle of the information. For instance, the @DrunkBuzzfeed account draws attention to BuzzFeed headlines and their form and content by mashing together parts of three random headlines). Commentary bots interact with the information beyond rebroadcasting or curating it, usually by augmenting it, providing an interpretation of it, or otherwise adding to the news content in some way.

Commentary Twitter bots emerge as an interesting category because they seem less prone to algorithmic discoverability—we sampled the DataSift dataset on a set of keywords derived from our initial purposive sample (which had 30 commentary bots—half of the sample), and got only five commentary bots in the larger sample. This may be in part due to the fact that they are less explicit about their bot nature in titles, bios, and handles, and in part due to the fact that their feed patterns and tweet content are not as formulaic or repetitive as those of topical news feeds or bots which aggregate and rebroadcast news and information. Due to the methodology we followed for collecting the purposive sample (i.e., looking for media reports of interesting bots), we must also consider the possibility that commentary bots were simply more likely to have gained some form of substantial external attention and therefore found their way more easily into our purposive sample.

The exact form of the output for topical, general, or niche feeds may vary, but tweets usually include story headlines and links to primary sources or aggregators, and may sometimes include other elements, like hashtags or images. Tweets may take the form of a data entry (with values or numbers) if the source of input is a database, and do not always include a link.

In the cases where the input comes from other tweets, whether standalone ones or replies/mentions of the bot, the output can take the form of a reply to the source tweet or a retweet of the source tweet, depending on the algorithm.

In terms of output type, the purposive sample was heavily skewed in favor of commentary accounts: 30 out of 60 bots were commentary ones (50 percent), whereas in the DataSift sample only 5 out of 178 were commentary accounts (2.8 percent). Instead, geo-specific news bots were represented much more extensively in the DataSift sample (43 out of 178, 24.1 percent) than in the purposive sample (3 out of 60, 5 percent). Other dominant categories that were similarly pervasive in both samples include topical feeds (16/60 in purposive sample, 26.7 percent; 60/178 in DataSift sample, 33.7 percent) and niche news bot accounts (22/60 in purposive sample, 36.7 percent; 70/178 in DataSift sample, 39.3 percent). The DataSift sample also had 14 non-news bot accounts, likely a result of our filtering keyword search catching the "botness" of the account despite its irrelevance in terms of news/information provision. This meant, for instance, that an account's bio contained the word "bot" or "automated," but the account itself tweeted personal reflections or spam links instead of what could be considered news or useful information. A detailed breakdown of output types is shown in Figure 2.

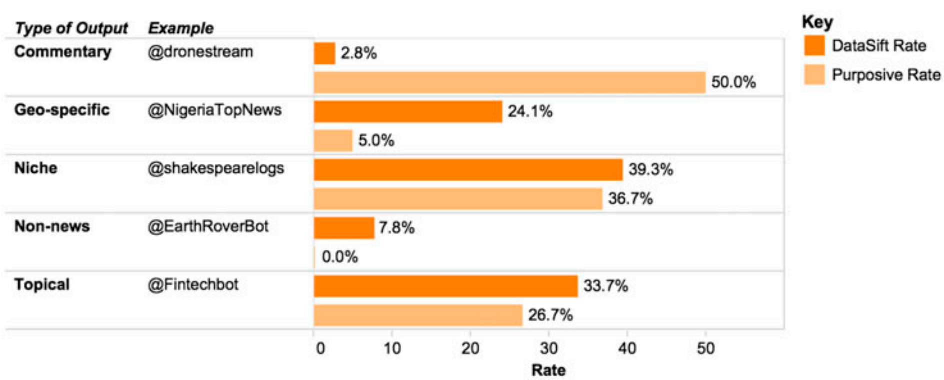

FIGURE 2
Prevalence of types of output (categories may overlap)

Algorithm

Another criterion along which the news bots can be classified is the nature of their driving algorithm: how does the bot turn the input into the output? One of the ways Twitter bots in our combined sample process the inputs into the outputs is *rebroadcasting* information onto the Twitter feed from a single source. *Rebroadcasting* can also be thought of as *bridge dissemination*, wherein the algorithm connects another platform to Twitter, creating a new outlet for information that previously existed only on that platform, i.e., Reddit or Instagram (as is the case with @rMLS_bot that tweets updates about major league soccer from the subreddit r/MLS).

Another type of algorithm is *aggregating/curating* information from multiple sources, either by way of search queries with a keyword or set of keywords or by using a pre-set list of sources (e.g., @SocialinMemphis aggregates news stories about the Memphis, TN area from various websites). Another subset of bots *react or respond* to information or data found or tweeted at them: @BBCWeatherBot responds to tweets with locations, while @ajamenergybot reacts to tweets from a pre-set list of congressional Twitter accounts that contain certain relevant keywords by retweeting them.

Augmenting information from a source or sources by adding data to it is another type of news bot algorithm (@NYTPlusContext adds links to quotes posted by @NYTMinusContext, and @cybercyber augments article links with a list of words containing "cyber" found in said articles). Some bots perform *data analysis* or *processing* of data from the input source, like @Treasury_io, which processes data from a US Treasury database and turns useful bits of the daily reports into tweets. Finally, some bots work by *generating* new data or information from an input source: e.g., @DrunkBuzzfeed mashes together three BuzzFeed headlines to create new headlines.

Our purposive sample had a more varied selection of types of algorithms than the DataSift sample, although in both cases curating/aggregating dominated in the samples, with 40 percent (24) and 44.4 percent (79) of accounts, respectively. But while the DataSift sample also had 43.3 percent of accounts (77) engaged in rebroadcasting/ bridging, the purposive sample had just 11.6 percent of accounts (7) using that

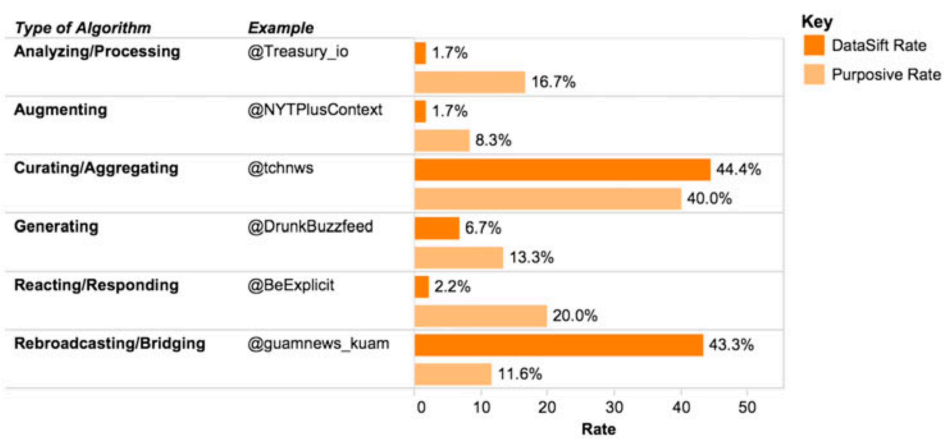

FIGURE 3
Prevalence of types of news bot algorithms (categories may overlap)

algorithm. On the other hand, analysis/processing and reacting/responding were present in the purposive sample to a much greater extent than in the DataSift sample (see Figure 3 for breakdown). This seems to align with the greater presence of commentary bots in the purposive sample, as they are more prone to using the more complex kinds of algorithms that engage with data or with users in more creative ways in their activity.

Intent/function

While the algorithm of a news bot (the "how" of its operation) is a convenient vehicle of typologizing, the intent or function of the bot (the "why" of its existence) is also an interesting dimension. Beyond simply *informing* the reader, news bots can have more complex functions, such as: *reporting/recommending breaking news* (@WikiLive-Mon, which draws on Wikipedia edits, recommends "breaking news candidates" based on the frequency of article edits in a given time period).

Some news bots are built to enable *discovery and investigation* of information that is hidden or hard to obtain (@Treasury_io reveals US treasury data from badly formatted reports, @stopandfrisk parses a 2011 database of incidents to bring attention to the scope of the problem). Yet other news bots are intended to replicate the journalistic function of ensuring *accountability* of those in power (@yourrepsonguns retweets congressional tweets mentioning firearms) or the media (@NYTAnon highlights every use of anonymous sources in *New York Times* articles).

A number of bot accounts provide *critique or opinion* on salient issues (@cybercyber, for instance, critiques the overuse of "cyber" in news stories) or on news values/content (@speak4yourself invites news media to not extrapolate in headlines, @BeExplicit combats passive voice use in news tweets, and @NYTAnon comments on using anonymous sources in news stories). Finally, some news bots are also built to provide a *service* in addition to relaying information (most common examples include traffic bots, weather bots, and currency exchange rate bots like @BTCtoUSD) and for *entertainment* purposes (e.g., @mot_namdeirf remixes sentences from a corpus of Tom Friedman *New York Times* columns).

The informing function dominated both the purposive sample and the DataSift sample (46.7 percent (28) and 82.6 percent (147), respectively), which is perhaps unsurprising as informing is a central goal of news (Kovach and Rosenstiel 2007). But overall, the purposive sample had more variety in the bots' intended functions than the DataSift sample (a detailed breakdown is shown in Figure 4). This can be linked to the greater presence of commentary bots in the purposive sample, as their creators seem to be more inventive in terms of the intent behind their creations, also enabled in part by the greater variety in the algorithms behind the bots.

Interestingly, in the DataSift sample there were several bots that were made to look like news feeds but were actually *spam* bots, with links in tweets leading to ad-infested intermediary pages. This might indicate that news bots are perceived as valuable by spammers, but also has implications for the transparency and credibility of news bots, which we address further in our discussion section.

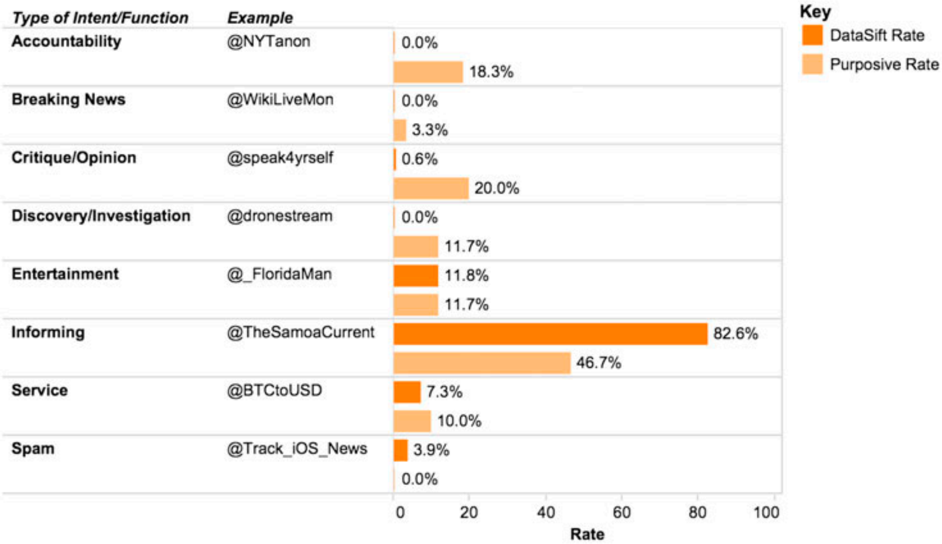

FIGURE 4
Prevalence of types of news bot intents/functions (categories may overlap)

Limitations

This study uses a targeted sampling strategy, working to build a purposive sample of known news bots and using it to inform the automated data collection for the larger DataSift sample. Our combined samples of course represent only a small fraction of overall news bot activity on Twitter and, furthermore, exclude bots that may be effectively hiding their "botness." Each sample has its limitations and biases and we make no claim that these non-random samples are representative of the overall news bot prevalence on the platform. Yet by contrasting the two samples, we are able to offer findings that are more robust and provide a broader picture of news bot activity. Our sampling strategy enabled the identification of accounts for a careful manual examination and open coding that offers an initial overview as well as rich, specific details into a diverse swath of news bot activity.

Since our study was observational in nature, we were somewhat limited in our ability to understand the intent of the creators of news bots in our sample. A further avenue of research seeking to more deeply examine the intent and motives behind news bot creation would require engaging with the creators/makers and interviewing them about their editorial and design decisions.

Another limitation of our study was our deliberate decision to limit our data collection to the English-language Twitter sphere. Other languages on Twitter and other social media may present their own interesting data and possible findings on the use of news bots (Abokhodair, Yoo, and Mcdonald 2015), but those are beyond the scope of our study, which concentrates on the English-language media and social networks.

Additionally, we only examined one social media platform, Twitter, due largely to the fact that the bot phenomenon on Twitter is well known and prevalent. This allowed us to have a solid grounding in the existing literature on bots and social automation as we examined a relatively new and under-researched niche of news bots.

Discussion

The use of "news bots"—automated accounts that participate in news and information dissemination on social networks—presents an intriguing opportunity for news organizations and journalists as they adapt to work in a dynamic digital media environment. Our study demonstrates that news bots are already being used in many creative ways by media organizations and individuals, but also raises some questions about the limits of automation in the journalistic workflow. The study also contributes to the more academic discussion of the role of algorithms in journalistic practices, and helps reconceptualize how scholars might approach editorial processes and newsroom decision-making in light of how automation changes, enables, and becomes integrated into various functions and strategies in the newsroom.

Our findings indicate that many news bots are used in very simple ways to rebroadcast or bridge existing traditional news platform content to social media, as well as to curate and aggregate content from multiple sources. Some of these activities have been reported in prior targeted studies of Twitter bots (Mittal and Kumaraguru 2014; Starbird et al. 2010). A few of the more interesting applications of the curation and filtering potential of news bots emerge in the sphere of niche news accounts and geo-specific news bots, especially local and hyperlocal ones. Niche and geo-specific bots create new opportunities to serve the needs of new, smaller audience groups and to fulfill their desire to aggregate news and information around narrow domains cheaply and at minimal marginal cost (Cohen, Hamilton, and Turner 2011). Given the prevalence of niche and geo-specific bots that we found, we expect that such bots could be strategically employed by media outlets or citizen journalists seeking to serve the information needs of micro-audiences.

Transparency of automated news and information services on social media emerges as another important issue with respect to the algorithmic accountability (Diakopoulos 2015) of the bots themselves and their creators and maintainers. What must we know about how bots work in order to trust them? Our findings indicate that not all news bots in our sample are transparent about their sources, about the algorithms behind their outputs, and even about the fact that they are bots. In fact, in our combined samples information sources used by bots were undefinitive in 45 percent of cases. This raises the question of how the phenomenon of journalistic transparency manifests in automated news and information production, and how journalists and news consumers should adapt their understanding of credibility and trustworthiness with respect to automation. Accountability also means that some entity is held accountable in case of a legal violation or in case of defamation or libel. This entity is likely to be the human creator or the organization of the creator for automated bot accounts, however, this may vary by jurisdiction. These considerations will gain importance as more news and information services are delivered by bots, demanding new journalistic norms be developed to account for these new scenarios (McBride and Rosenstiel 2013).

Commentary news bots that augment content or deliver critique and opinion raise the issue of the limits of automated commentary on privately owned social media platforms in terms of third-party ability to impose their own rules and terms on those who use their website. Twitter estimates that about 8.5 percent of the accounts on the platform are automated in some way, and it routinely scrubs out flagrant spam bots.

But the same terms of use that Twitter uses to expunge those spam accounts might also be brought to bear on a more noble news bot. If Twitter believes a news bot is violating its terms of service, whether because of its algorithm (@BechdelBot was briefly blocked from using the Twitter API [application programming interface] because of a high number of requests to the database) or because of the nature of the opinions the bot espouses, the corporate platform owner can easily decide to shut down the account. Moreover, individuals who disagree with the output or intent of the bot could flag it as "offensive" or "spam" and similarly trigger platform censorship of the bot. The broader issue really gets at the limits of freedom of speech on such platforms, and the degree of freedom of speech that should be afforded to news bots, which are arguably operating under human communicative intentions and designs. Who will decide what rights news bots have as opposed to humans doing journalism and critiquing those in power?

Napoli (2015) raises an important point about the extent to which social media platforms and the algorithms behind them reflect public interest values. In our case, this is complicated by the idea that newsrooms, guided by journalistic norms, can exert an editorial function on the existing third-party platforms, by building and designing algorithms that enable an "individualist model of public interest," where users are empowered to make decisions in relation to their own news and information consumption. How the algorithms (e.g., commentary bots) are designed to moderate and enable such empowerment deserves close scrutiny by both academics and journalism practitioners and should be part of the editorial design process.

It is also worthwhile to discuss the limitations of automated news commentary in terms of import and weight that such commentary may possess in comparison with traditional accountability journalism (Kovach and Rosenstiel 2007). Holding someone accountable through the news media acts by inducing public pressure that may demand a response from that party—but do bots deserve a response? In some cases, at least, it seems bots that provide critique are able to start a constructive discussion: the @NYTAnon account generated a robust and lengthy online discussion about the practice of using anonymous sources in mainstream US media. But whether any policy or editorial changes will follow remains to be seen.

As automated news and information production develops, our study reveals some avenues for future research and analysis in this area. One such issue is the intentionality of news bots and algorithms: the creator's intent and its interpretation by the algorithm will contribute to the discussion about transparency and credibility of automated journalism. Another important question for newsrooms will be to understand how a media outlet's editorial criteria can manifest in algorithmic form, and what sorts of new decisions editors will have to begin making with regards to designing, building, editing, managing, and decommissioning news bots. Gillespie (2014) calls these kinds of editorial automation practices "public relevance algorithms" and suggests both researchers and media practitioners should interrogate these as a recent but already key feature of the news and information ecosystem, especially with regard to how such algorithms evaluate information relevance and inclusion, and how they manage the notion of objectivity. Finally, the issue of information verification, which has risen on the media agenda with the proliferation of social media platforms and citizen-generated content (Castillo, Mendoza, and Poblete 2011), will also be salient for newsroom bot designers. It remains to be seen if a news bot can be designed to perform the information gathering, fact-checking, and content verification functions that human journalists now perform. How robust can a news bot's design be in the face of data quality issues or human errors that they might otherwise unwittingly re-transmit? As Carlson (2015) notes, the emerging practice of automated reporting presents a challenge in how it might alter the working practices of journalists and newsrooms, but also affects larger understandings of what journalism is and how it ought to operate—questions which undoubtedly concern academics who study the theory and practice of journalism.

Conclusions

In this study, we analyzed a sample of news bot accounts on Twitter to understand how these bots are currently being used and to examine how using automation and algorithms may change the modern media environment. Based on our analysis and the news bot typology that emerged from it, we articulate a design and editorial decision space intended to guide designers in defining the intent, utility, and functionality of future news bots.

The design and editorial decision space is informed by four key categories that emerged during the open coding and typologizing process. These categories reflect both the bots' algorithmic logic and the editorial decisions behind it. The categories include the inputs and sources of input data; the outputs produced by the news bots; the algorithms that guide how a news bot turns inputs into outputs; and the function or intent of the news bot. These categories and their subcategories can act as a starting point for the decisions a newsroom team might have to make in defining and building a bot.

The proposed design space highlights the limits of news bots (e.g., automated commentary and opinion, algorithmic transparency and accountability), notable lacunas and opportunities for designing solutions (fact-checking and verification bots, question-asking and reporting bots, interactive news), and areas where news bots may enable innovation, such as niche and geo-specific news and information. We emphasize these in our discussion and examine the challenges and opportunities to using news bots as instruments of journalism.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the authors.

FUNDING

No grants were used to support this research.

SUPPLEMENTAL DATA

The raw data, including the accounts that were analyzed for this study, have been made available for research purposes and can be downloaded at https://goo.gl/ORU1Yz (Google spreadsheet, can be downloaded as .xls or .csv).

NOTES

1. See http://blog.automatedinsights.com/post/109491692518/automation-helps-ap-publish-10-times-more-earnings.
2. See http://www.newrepublic.com/article/121551/bot-bubble-click-farms-have-inflated-social-media-currency.
3. See http://tinysubversions.com/botsummit/2014/.
4. See http://www.bbc.com/news/technology-16108876.
5. See http://www.technologyreview.com/news/428286/twitter-mischief-plagues-mexicos-election/.
6. See http://truthy.indiana.edu/botornot/.

REFERENCES

Abokhodair, Norah, Daisy Yoo, and David W. Mcdonald. 2015. "Dissecting a Social Botnet." In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing – CSCW '15*, 839–851. New York: ACM.

Aiello, Luca Maria, Martina Deplano, Rossano Schifanella, and Giancarlo Ruffo. 2012. "People are Strange when you're a Stranger: Impact and Influence of Bots on Social Networks." *Links* 697 (483, 151): 1–566.

Anthony, Laurence. 2004. "AntConc: A Learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit." In *Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning*, 7–13. Tokyo: Waseda University.

Barbour, Rosaline S. 2001. "Checklists for Improving Rigour in Qualitative Research: A Case of the Tail Wagging the Dog?" *British Medical Journal* 322 (7294): 1115–1117.

Boshmaf, Yazan, Ildar Muslukhov, Konstantin Beznosov, and Matei Ripeanu. 2011. "The Socialbot Network: When Bots Socialize for Fame and Money." In *Proceedings of the 27th Annual Computer Security Applications Conference*, 93–102. New York: ACM.

Broussard, Meredith. 2014. "Artificial Intelligence for Investigative Reporting: Using an Expert System to Enhance Journalists' Ability to Discover Original Public Affairs Stories." *Digital Journalism*. doi:10.1080/21670811.2014.985497.

Carlson, Matt. 2015. "The Robotic Reporter: Automated Journalism and the Redefinition of Labor, Compositional Forms, and Journalistic Authority." *Digital Journalism* 3 (3): 416–431.

Castillo, Carlos, Marcelo Mendoza, and Barbara Poblete. 2011. "Information Credibility on Twitter." In *Proceedings of the 20th International Conference on World Wide Web*, 675–684. New York: ACM.

Chu, Zi, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. 2010. "Who is Tweeting on Twitter: Human, Bot, or Cyborg?" In *Proceedings of the 26th Annual Computer Security Applications Conference*, 21–30. New York: ACM.

Cohen, Sarah, James T. Hamilton, and Fred Turner. 2011. "Computational Journalism: How Computer Scientists Can Empower Journalists, Democracy's Watchdogs, in the Production of News in the Public Interest." *Communications of the ACM* 54 (10): 66–71.

Cook, David M., Benjamin Waugh, Maldini Abdipanah, Omid Hashemi, and Shaquille Abdul Rahman. 2014. "Twitter Deception and Influence: Issues of Identity, Slacktivism, and Puppetry." *Journal of Information Warfare* 13 (1): 58–71.

Diakopoulos, Nicholas. 2015. "Algorithmic Accountability: Journalistic Investigation of Computational Power Structures." *Digital Journalism* 3 (3): 398–415.

Dickerson, John P., Vadim Kagan, and V. S. Subrahmanian. 2014. "Using Sentiment to Detect Bots on Twitter: Are Humans More Opinionated Than Bots?" In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference*, 620–627. NewYork: ACM.

Edwards, Chad, Autumn Edwards, Patric R. Spence, and Ashleigh K. Shelton. 2014. "Is that a Bot Running the Social Media Feed? Testing the Differences in Perceptions of Communication Quality for a Human Agent and a Bot Agent on Twitter." *Computers in Human Behavior* 33: 372–376.

Ferrara, Emilio, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2014. "The Rise of Social Bots." *arXiv Preprint arXiv:1407.5225*.

Gao, Tong, Jessica R. Hullman, Eytan Adar, Brent Hecht, and Nicholas Diakopoulos. 2014. "NewsViews: An Automated Pipeline for Creating Custom Geovisualizations for News." In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, 3005–3014. New York: ACM.

Gillespie, Tarleton. 2014. "The Relevance of Algorithms." In *Media Technologies: Essays on Communication, Materiality, and Society*, edited by Tarleton Gillespie, Pablo Boczkowski, and Kirsten Foot, 167–193. Cambridge, MA: MIT Press.

Hall, Stuart. 1981. "The Determination of News Photographs." In *The Manufacture of News: A Reader*, edited by Stanley Cohen and Jock Young, 226–243. Beverly Hills, CA: Sage.

Haustein, Stefanie, Timothy D. Bowman, Kim Holmberg, Andrew Tsou, Cassidy R. Sugimoto, and Vincent Larivière. 2015. "Tweets as Impact Indicators: Examining the Implications of Automated Bot Accounts on Twitter." *Journal of the Association for Information Science and Technology*. arXiv:1410.4139v1.

Hwang, Tim, Ian Pearce, and Max Nanis. 2012. "Socialbots: Voices From the Fronts." *Interactions* 19 (2): 38–45.

Jupp, Victor. 2006. *The SAGE Dictionary of Social Research Methods*. Sage.

Kovach, Bill, and Tom Rosenstiel. 2007. *The Elements of Journalism: What Newspeople Should Know and the Public Should Expect*. New York: Three Rivers Press.

Larsson, Anders Olof, and Hallvard Moe. 2014. "Bots or Journalists? News Sharing on Twitter." *Communications – The European Journal of Communication Research*. http://www.academia.edu/9372564/Bots_or_Journalists_News_Sharing_on_Twitter

Lee, Kyumin, Brian David Eoff, and James Caverlee. 2011. "Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter." In *Proceedings of the 5th AAAI International Conference on Weblogs and Social media*. Menlo Park, CA: The AAAI Press.

Lofland, John, and Lyn H. Lofland. 2005. *Analyzing Social Settings: a Guide to Qualitative Observation and Analysis*. Belmont, CA: Wadsworth.

McBride, Kelly, and Tom Rosenstiel, eds. 2013. *The New Ethics of Journalism: Principles for the 21st Century*. Thousand Oaks, CA: CQ Press.

Messias, Johnnatan, Lucas Schmidt, Ricardo Oliveira, and Fabrício Benevenuto. 2013. "You Followed my Bot! Transforming Robots into Influential Users in Twitter." *First Monday* 18 (7).

Mittal, Sudip, and Ponnurangam Kumaraguru. 2014. "Broker Bots: Analyzing Automated Activity During High Impact Events on Twitter." *arXiv preprint arXiv:1406.4286*.

Mustafaraj, Eni, and P.Takis Metaxas. 2010. “From Obscurity to Prominence in Minutes: Political Speech and Real-Time Search.” In *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line*. Raleigh, NC: US.

Napoli, Philip M. 2015. “Social Media and the Public Interest: Governance of News Platforms in the Realm of Individual and Algorithmic Gatekeepers.” *Telecommunications Policy*.

Ratkiewicz, Jacob, Michael Conover, Mark Meiss, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. 2011. “Detecting and Tracking Political Abuse in Social Media.” In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 297–304. Menlo Park, CA: The AAAI Press.

Shearer, Matt, Simon Basile & Geiger, Clement. 2014. “Datastringer: Easy Dataset Monitoring for Journalists.” In *Proceedings of Symposium on Computation + Journalism*.

Shoemaker, Pamela J., Tsan-Kuo Chang, and Nancy Brendlinger. 1987. “Deviance as a Predictor of Newsworthiness: Coverage of International Events in the US Media.” In *Communication Yearbook*, edited by Margaret L. McLaughlin, Vol. 10: 348–365. Beverly Hills, CA: Sage.

Starbird, Kate, Leysia Palen, Amanda L. Hughes, and Sarah Vieweg. 2010. “Chatter on the Red: What Hazards Threat Reveals About the Social Life of Microblogged Information.” In *Proceedings of the 2010 ACM conference on Computer Supported Cooperative Work*, 241–250. New York: ACM.

Steiner, Thomas. 2014. “Telling Breaking News Stories from Wikipedia with Social Multimedia: A Case Study of the 2014 Winter Olympics.” *arXiv preprint arXiv:1403.4289*.

Tavares, Gabriela, and Aldo Faisal. 2013. “Scaling-Laws of Human Broadcast Communication Enable Distinction between Human, Corporate and Robot Twitter Users.” *PLoS ONE* 8 (7): e65774.

Wagner, Claudia, Silvia Mitter, Christian Körner, and Markus Strohmaier. 2012. “When Social Bots Attack: Modeling Susceptibility of Users in Online Social Networks.” In *Proceedings of WWW: 21th International Conference on World Wide Web*, 41–48. New York: ACM.

Wald, Randall, Taghi M. Khoshgoftaar, Amri Napolitano, and Chris Sumner. 2013. “Predicting Susceptibility to Social Bots on Twitter.” In *Information Reuse and Integration (IRI), 2013 IEEE 14th International Conference*, 6–13. Piscataway, NJ: IEEE.

Wilkie, Alex, Mike Michael, and Matthew Plummer-Fernandez. 2014. “Speculative Method and Twitter: Bots, Energy and Three Conceptual Characters.” *The Sociological Review* 63 (1): 79–101.

Young, Mary Lynn, and Alfred Hermida. 2014. “From Mr and Mrs Outlier to Central Tendencies: Computational Journalism and Crime Reporting at the Los Angeles Times.” *Digital Journalism* 3 (3): 381–397.

Tetyana Lokot, (author to whom correspondence should be addressed), Philip Merrill College of Journalism, University of Maryland, USA. E-mail: tlokot@umd.edu. ORCID http://orcid.org/0000-0002-2488-4045

Nicholas Diakopoulos, Philip Merrill College of Journalism, University of Maryland, USA. E-mail: nad@umd.edu. ORCID http://orcid.org/0000-0001-5005-6123